DOCUMENT RESUME

ED 158 726 IR 006 181

AUTHOR	Bunderson, C. Victor; Schneider, Edward W.
TITLE	Formative Evaluation Fundamentals for TICCIT Courseware. Occasional Paper No. 2.
INSTITUTION	Brigham Young Univ., Provo, Utah. Inst. for Computer Uses in Education.
SPONS AGENCY	Mitre Corp., McLean, Va.; National Science Foundation, Washington, D.C.
PUB DATE	Feb 74
CONTRACT	C-179
NOTE	18p.; Institute for Computer Uses in Education Series; For related documents, see IR 006 174, 177, 178 and IR 006 180, 182, 183
EDRS PRICE	MF-$0.83 HC-$1.67 Plus Postage.
DESCRIPTORS	*Community Colleges; *Computer Assisted Instruction; *Course Evaluation; Display Systems; *Evaluation Methods; *Formative Evaluation; Guidelines; Higher Education
IDENTIFIERS	Educational Testing Service; *TICCIT Computer System

ABSTRACT

The formative evaluation process described was designed to be used by Brigham Young University personnel in 1974 to evaluate and improve TICCIT programs and materials for community college credit courses. Involving the use of editorial judgment and data on student use, this process includes five levels of evaluation and revision: (0) materials are reviewed for subject matter accuracy, instructional psychology, and message design; (1) materials pass through several cycles of formal debug, using skilled and critical students to look at every display with different types of mental sets characteristic of students and catch problems in answer processing; (2) lessons and units are tested on 20 or more students enrolled in convenient institutions to identify those lessons, segments, and displays which produce confusion or difficulty; (3) courses, lessons, and units are tested on 20 or more community college students to discover any remaining areas causing difficulties; and (4) courses, units, and lessons are tested on several hundred community college students. Modifications indicated by the results in each step are part of the on-going process. The most advantageous role during this period for Educational Testing Service (ETS), which would perform a separate summative evaluation of the final revised version, was seen as facilitator of interchange between the developers and the users and as a facilitator of the formative evaluation process. (Author/CMV)

INSTITUTE FOR COMPUTER USES IN EDUCATION

FORMATIVE EVALUATION FUNDAMENTALS FOR TICCIT COURSEWARE

C. Victor Bunderson
Edward W. Schneider

Occasional Paper No. 2

February 1974

Sponsored by:
The MITRE Corporation
under NSF Contract #C-179

DIVISION OF INSTRUCTIONAL RESEARCH, DEVELOPMENT & EVALUATION

Brigham Young University

Provo, Utah

FORMATIVE EVALUATION FUNDAMENTALS FOR TICCIT COURSEWARE

C. Victor Bunderson
Edward W. Schneider
Brigham Young University
February 1974

This paper is designed to communicate certain concepts regarding formative evaluation of TICCIT courseware to teachers and administrators who plan to use the TICCIT system.

The terms "formative evaluation" and "summative evaluation" were introduced and distinguished by Michael Scriven in 1967 (Scriven, 1967). Bloom (1956), has written a handbook for formative and summative evaluation for teachers which is extremely useful as a guide for teachers involved in small instructional development and evaluation projects. In general, the distinction useful to the TICCIT project is that formative evaluation, performed by Brigham Young University personnel, will exercise editorial judgment, and collect data on student use which can be used to improve the courseware until the final revised version is installed in the summer of 1975. Summative evaluation, on the other hand, is something which Educational Testing Service (ETS) will perform independently, providing an overall evaluation of the final system. The system will be compared with its design goals and with other important instructional objectives.

The scientific method has been defined as "doing your damnedest with your mind with no holds barred" to push forward the purposes of science. A working definition of formative evaluation might be: formative evaluation is a process of doing your damnedest with human judgment and student data to locate and improve deficiencies in content accuracy, instructional effectiveness, sensible and responsive decision logic, and the organic unity and esthetics of instructional material.

Since both human judgment and data are used to improve various aspects of the courseware, instead of just talking about a "formative evaluation process," it is useful to distinguish five levels of evaluation and revision through which courseware will pass. These five levels are as follows:

0) Lessons (and unit material) are reviewed and revised for subject matter accuracy and excellence, instructional psychology, and message design. Lessons are input and mechanics debug corrects "proofreading" level errors in the displays and logic.

1) Lessons (and unit material) pass through several cycles of formal debug, using skilled and critical students who look at every display with different types of mental sets characteristic of students. The majority of problems in answer processing will be caught and corrected by this step. Courseware is now in a form acceptable to be used by students in credit classes, but these classes should be backed up by more teaching personnel than will be necessary later, for some percentage of the segments will fail to teach adequately.

2) Lessons and units will be tested on 20 or more students enrolled in convenient institutions (Utah Valley). Statistics will be used to identify those lessons, segments, and displays which produce confusion or difficulty. These will be revised.

3) Courses, lessons, and units will be tested on 20 or more community college students at PC and NVCC. Based on these data, courses,

units, lessons, and displays which produce confusion or difficulty will be modified. Aspects of the system design may be modified. Aspects of the implementation plan and faculty role models which cause problems or fall short will be modified.

4) Courses, units, and lessons will be tested on several hundred community college students at PC and NVCC. The same actions as in stage 3 will be taken as indicated by the data.

Table 1 relates the five levels above to the definition of formative evaluation. Recall that this definition dealt with the application of human judgment and student data to improvements in content, effectiveness, message design, decision logic and other matters. In Table 1 the five levels are listed as subheadings and there are two columns, one dealing with the application of human judgment and the other dealing with the application of student data to the process of making revisions. You will note that revisions in content always rely on human judgment, in this case the judgment of subject-matter experts. This judgment is applied first among the development group and later at the colleges as faculty members provide input on content in accuracies and questions. Instructional effectiveness may be addressed by the human judgment of an instructional psychologist at an early stage in development. Ultimately, however, effectiveness becomes a question which can only be answered on the basis of student data. Message design is mixed between judgment and data, as indicated by the horizontal brackets in Table 1. Human judgment is used through the manuscript level. Following that, student data are collected before any message design revisions are made.

Revising the decision logic is purely a function of student data. It is only through the experience of a number of students, working through the answer processing, instance files, and other complex parts of the instructional logic that we can find loops, blind alleys, incongruities, and other difficulties. It is usually a fairly slow process requiring many, many students to take the material. Data from their interactions must be recorded and summarized before all little problems in decision logic can be ironed out.

The effectiveness of instruction in segments and lessons is a matter which can be addressed by instructional psychologists at the manuscript level, but really can only be answered satisfactorily through the application of student data in levels 2, 3, and 4. It is an axiom in instructional psychology that human judgment should be used exclusively only when data are lacking. It is probably true that some psychologists have failed to use judgment at times, having become too cautious and distrustful of intuition, emotion, creativity, etc. Some of them may lose the broad prospective needed for good courseware development. A psychologist strictly from the laboratory tradition and without a certain feeling for art, style, and creativity is not always a good team member. We hope that we have been able to avoid this pitfall.

The five evaluation levels planned for TICCIT courseware should be compared with the evaluation and revision methods used in existing modes of instruction. A typical mode would be a lecture course, planned and executed by a single faculty member. Another would be the process of

developing a new textbook. These revealing comparisons are presented in Table 2. Table 2 is organized in three columns. In the first column we have described each of the five computer-assisted instruction evaluation levels. In column 2 we discuss the analogous procedure for the preparation of a lecture class, and in column 3 we compare the procedure for a textbook.

In viewing the analysis presented in Table 2 it can be seen that the quality control of level 1 TICCIT courseware is more than equivalent to a textbook or to a new lecture course introduced by a faculty member. The editorial review from the perspectives of subject matter, instructional psychology, and message design as well as the usual proofreading will make the courseware not only respectable, but roughly equal to the presently utilized "hard-copy" materials. The extent to which this individualized material will effectively teach different kinds of students is still an issue, but it cannot be resolved until data are collected in connection with steps 2, 3, and finally step 4. Obviously, step 2, which uses BYU-students who are not drawn from the population of community college students, will be less accurate in identifying particular instructional weaknesses in lessons and segments than data collected at the colleges themselves.

If the colleges feel that level 1 courseware would be suitable for administration to their students (as we feel it would be for our BYU students), then the only questions are:

1) How to back the material up with support from teachers sufficient to assure that the students learn well?

2) How to maintain adequate control of data collection so that the data will be correctly interpreted?

The first question can best be answered with data which give us some idea of the percentage of lessons which might be difficult or ambiguous. If this number is relatively small (say less than 10 percent), it would seem that the risk would be small for going ahead within the colleges with several formative evaluation classes. If the number is larger (say 30 percent), then there may be a question as to whether it would be appropriate to do this. There is also a question whether we will have sufficient resources left to revise 30 percent or more of an already very extensive body of material. It is most unfortunate that it has been impossible to test sample lessons on the TICCIT system until this late in the project. A great loss of time and money occurs because of the lack of information about how students will respond to the final product. An enormous amount of revision effort could be avoided by the ability to test developing lessons with students.

If the colleges decide to conduct a level 3 evaluation on site starting September, 1974, then they may obtain benefits in terms of more effective courseware and in terms of faculty development. The process of working at either level 3 or level 4 in assisting in the collection of formative evaluation data would be a new experience for most teachers and could be a challenging and interesting experience. Never before has a community college had the ability to get such closely detailed data, scrutinizing each lesson, segment, and individual frame. Should teachers obtain some kind of professional development credit for working in this environment? Could this type of work lead to the creation of development expertise at the colleges? In other words, can developmental evaluation teach enough

about the various courseware components and how students respond to them to give faculty members an intuitive sense for the structure and function of learner controlled courseware? Would this enable them to learn rapidly how to develop it themselves? These are some questions which must be answered in developing a plan for implementation and testing next fall.

The probability that MITRE and BYU will complete very little courseware through level 2 before installation in colleges raises questions about the best role which Educational Testing Service should play. It does not seem to serve the best interests of the colleges, BYU, MITRE, or the field of computer-assisted instruction to view the year 1974-75 as a summative evaluation year. Because of schedule slips, the courseware will not have a fair chance to go through the formative evaluation process for effectiveness of lessons and segments. Because of the expertise ETS has in test and instrument design, and because of the good association they have with both colleges, ETS might best function during the year 1974-75 by facilitating the interchange between the developers and the users and in facilitating the formative evaluation process which will be going on that year.

TABLE 1

USE OF HUMAN JUDGMENT AND STUDENT DATA

IN FIVE STEPS OF FORMATIVE EVALUATION

	Human Judgment		Student Data		
Level	Content	Instructional Effectiveness	Message Design	Decision Logic	Instructional Effectiveness
0 Expert Reviewers	Independent Review	Instructional Psychologist Review	Message Design Expert		
1 Bright Students	Bright Student	Bright Student	X	X	X
2 20-30 Students at BYU	*	*	XX	XX	XX
3 20-30 Students at College(s)	A Few College Faculty	*	* XXX	XXX	XXX
4 Hundreds of Students at College(s)	Numerous College Faculty	*	* XXXX	XXXX	XXXX

X The number of X's indicate relative importance of different sources of student data in indicating needed revisions.

* While human judgment is obviously not abandoned in these cases, where student data are available, human judgment (which often reflects personal taste and style) must take full account of it before revisions are made.

TABLE 2

A COMPARISON BETWEEN COMPUTER-ASSISTED-INSTRUCTION, LECTURES AND TEXTBOOKS IN REGARD TO STAGES OF FORMATIVE EVALUATION AND REVISION

CAI	LECTURES	TEXTBOOK
Level 0. Editorial evaluation by subject matter experts, instructional psychologists and message design experts. Revision of manuscript material.	A new set of lecture notes are rarely reviewed by an independent subject matter expert and never by instructional psychologists and message design experts. Handouts prepared to accompany the lecture may be proofread by a secretary.	Textbooks are independently reviewed by subject matter experts for content, prose and layout. Message design expertise is used only in a narrow sense, constrained by traditional page layout formats.
Level 1. Three or four expert students go through a detailed formal debug procedure. Errors, especially in logic and answer processing, are corrected.	Faculty members rarely, if ever, take time to subject their lecture notes to three or four critical students and revise them accordingly before they first deliver the course.	The manuscripts for textbooks are on some occasions exposed to some of the author's better students who will read and make appropriate comments.

Table 2 Continued

CAI	LECTURES	TEXTBOOK
Level 2. Material which has been debugged formally is exposed to 20 or 30 students at the development site (BYU in this case). Data are collected and revisions are made.	Not applicable because class lectures are rarely designed for transportability.	An author may expose the manuscript material to classes of students and test it informally at his home institution. There are, however, no formal procedures for collecting data and focusing these data on specific lessons, segments and individual frames as will be done by the TICCIT data reduction system. The possibility for collecting data this detailed is simply not available to the author of a manuscript.
Level 3. TICCIT courseware is exposed to classes of 20-30 students in sections at the colleges where it will be installed. Data are collected and sent back to BYU for the appropriate revisions.	Usually no data are collected, but revisions are made based on a teacher's subjective interpretation of students' reactions and complaints during the first semester. He will usually make revisions based on his anecdotal information, plus his own feelings.	Textbooks are rarely tested at other colleges, although some professors have colleagues at another university who are willing to try their manuscript in a class. Feedback for revision is quite informal and subjective.

Table 2 Continued

CAI	LECTURES	TEXTBOOK
Level 4. Based on data from hundreds of students, developers revise lessons and segments that do not help enough students to succeed on lesson tests.	Over a long period of use the lecture notes for a given faculty member's course are revised based upon the teacher's subjective interpretation of students' reactions. He does no formal data collection to focus his revision efforts; this is not a true formative evaluation.	Data are rarely collected, but users may send back comments on typographical errors and other matters. These are corrected in succeeding editions of the book. The author may revise the book in four or five years, but he does this to correct and update the content, rather than to improve the instructional effectiveness of the textbook. At least one calculus text has been corrected by offering a $5 reward for each new error detected in the practice problem solutions.

REFERENCES

Bloom, B. S. (Ed.) "Taxonomy of Educational Objectives: The Classification of Educational Goals." Handbook 1. Cognitive Domain. New York: McKay, 1956.

Scriven, M. "The Methodology of Evaluation." AERA Monograph Series on Curriculum Evaluation, 1967, No. 1, pp. 39-83.